AD__________________

Award Number: DAMD17-00-1-0613

TITLE: Validation of Causal Analysis for Obtaining Intervention-Study Results from Non-Intervention Studies

PRINCIPAL INVESTIGATOR: Mikel G. Aickin, Ph.D.

CONTRACTING ORGANIZATION: Kaiser Foundation Research Institute
Oakland, California 94612

REPORT DATE: October 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20020206 122

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
October 2001

3. REPORT TYPE AND DATES COVERED
Final (25 Sep 00 - 24 Sep 01)

4. TITLE AND SUBTITLE
Validation of Casual Analysis for Obtaining Intervention-Study Results form Non-Intervention Studies

5. FUNDING NUMBERS
DAMD17-00-1-0613

6. AUTHOR(S)
Mikel G. Aickin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Kaiser Foundation Research Institute
Oakland, California 94612

E-Mail: Mikel.Aickin@kpchr.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Recent research on breast cancer etiology has two important characteristics. (1) Most studies focus on a single potential causal agent, and (2) a large number of such agents have been studied, mostly retrospectively. This project performed a literature search in order to categorize recent studies with respect to their inferential structure, and risk factors investigated. Based on this review, this project developed new methods in time-to-event analysis that support a simulation/causation approach to the study of breast cancer. A new method of representing time-to-events was developed. It shows that the Kaplan-Meier method is appropriate for the simulation/causation approach, but that it cannot be used in retrospective studies. The bias was computed explicitly, and a new complementary exponential method for unbiased estimation of incidence rates in retrospective studies was developed. Although the methods of modern causal analysis can be extended to retrospective studies, their atemporal nature makes them less useful from a simulation/causation perspective.

14. SUBJECT TERMS
Breast cancer etiology, causal analysis

15. NUMBER OF PAGES
16

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

Table of Contents

Introduction

The specific aim of this project was to explore the extent to which one might employ data from existing studies to deduce the results of hypothetical intervention studies, using the concept of causation espoused by current researchers (primarily Judea Pearl, Clark Glymour, Richard Scheines, and Peter Spirtes, and others).

The premise of the aim was that although a causal analysis of "observational" studies always requires a certain amount of faith, expressed as analytic assumptions, it is better to try to use the existing data for some kind of causal analysis with prudent assumptions than it is to allow the data to be discarded. A further underlying premise was that causal analysis of breast cancer etiologic data would not be found in the literature.

Body

Methods

The methods were based on literature search and theoretical considerations. Current databases were searched for current and past trials. From this it became obvious that a Medline search was necessary. The theoretical considerations followed from the literature search.

Literature Search

Prospective trials in breast cancer were primarily treatment trials. Among the remainder, most were either inaccessible (because they are not completed), or tertiary prevention drug trials (such as Tamoxifen). It seems fairly clear that results from large trials like WHI, WHEL, WINS, and so on, will not become generally available for a considerable time yet. This suggested focusing the specific aim on the existing literature, and for this it was necessary to characterize that literature, with respect to analytic methods.

A Medline search of the recent literature used the search string

breast neoplasms/epidemiology[mh] AND risk factors[mh] NOT therapeutics[mh]

It found 764 papers 1993-present, of which 316 contained sufficient data or data references to be classified. The unclassified papers consisted primarily of editorials, or of reviews (meta-analyses, overviews, evidence-based reviews), secondarily of studies that were focused on specific biochemical mechanisms, but with no data or no human data. The search was not intended to be exhaustive, but to be representative of the breast cancer etiology literature.

Literature Results

This section summarizes some of the findings, primarily to indicate the type of data that was generated.

TABLE 1.

```
Prospective or |           Incidence or Mortality Study
Retrospective  |   Neither       Incid        Mort        Both  |     Total
---------------+----------------------------------------------+----------
       Neither |        32           5           0           0  |        37
         Prosp |         2          92          22           3  |       119
         Retro |         4         154           0           1  |       159
          Both |         0           0           0           1  |         1
---------------+----------------------------------------------+----------
         Total |        38         251          22           5  |       316
```

Table 1 shows that retrospective studies outnumber prospective studies by about 4:3. This was surprising, in that this ratio was predicted to be much higher. There were no retrospective studies of mortality. Both prospective and retrospective studies were heavily weighted toward analysis of breast cancer incidence, as opposed to mortality.

In order to classify the structure of the inference in the articles that were found, a symbology for representing inference structures was developed. An exposition of this symbolic representation of analytic structures is not detailed here. Table 2., however, shows the number of articles in each of the 30 structures that were found.

TABLE 2.

```
  Structure   |      Freq.     Percent        Cum.
--------------+-----------------------------------
01: y<x       |        137       43.35       43.35
02: y<x&f     |         46       14.56       57.91
03: x<>x      |         25        7.91       65.82
04: y<mx      |         22        6.96       72.78
05: y|s<x     |         21        6.65       79.43
06: y<x|s     |         17        5.38       84.81
07: y<x|y     |         10        3.16       87.97
08: y<xf      |          5        1.58       89.56
09: y<mx&f    |          3        0.95       90.51
10: y<x<>x    |          3        0.95       91.46
11: y<mx|y    |          3        0.95       92.41
12: y|s<x|s   |          2        0.63       93.04
13: y<z       |          2        0.63       93.67
14: x<y       |          2        0.63       94.30
15: y<x&f|y   |          2        0.63       94.94
16: y|s<x&f   |          2        0.63       95.57
17: y<x&f|s   |          1        0.32       95.89
18: y<xf|s    |          1        0.32       96.20
19: x<x&f     |          1        0.32       96.52
20: y<xf|y    |          1        0.32       96.84
21: y<my      |          1        0.32       97.15
22: x         |          1        0.32       97.47
23: y|s<xf    |          1        0.32       97.78
24: x|s       |          1        0.32       98.10
25: y<x<>x|   |          1        0.32       98.42
26: y|s<x|y   |          1        0.32       98.73
27: x<y|s     |          1        0.32       99.05
28: y|x<x     |          1        0.32       99.37
29: ?         |          1        0.32       99.68
30: y<x>x|y   |          1        0.32      100.00
--------------+-----------------------------------
        Total |        316      100.00
```

Here, the number before the colon is the rank of the structure, and then the symbols indicate the most complex analysis that was presented in the article. (Obviously, ranks above 16 do not have meaning.) More than 43% of articles used the simplest possible inferential structure (y<x), meaning that a single risk factor x was related to either breast cancer incidence or mortality y. An additional 15% employed the second simplest structure (y<x&f), in which other factors (often unspecified) were included in a multi-explanatory model (such as logistic regression). The structures that showed any reasonable degree of complexity appeared in 50 (16%) of articles. These included y<mx (permitting an effect modifier m), y<x|s (examination of a single risk factor in several strata), y<xf (models with interaction terms), y<mx&f (including both effect modifiers and other factors linearly), y|s<x&f (subtypes of breast cancer related to a single risk factor and linearly adjusted for other factors), y|s<xf (the same, but with interactions).

It is to be emphasized that while 31 % of articles did something more complex than a simple y<x analysis, in no case were causal models fitted to the data. This was the case despite the fact that there is no obvious reason why prospective studies cannot be analyzed using causal methods, and that the discovery of causal pathways and indirect causal influences would seem to be of some use in understanding the entire disease etiology. The failure to analyze retrospective studies causally is perhaps explained by the fact that the issues here have not been worked out in the literature, a point that will be discussed below.

TABLE 3.

```
                            Risk Factor |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
   01:reprod hist                       |         50        8.53        8.53
   02:family hist                       |         38        6.48       15.02
   03:age                               |         30        5.12       20.14
   04:alcohol                           |         23        3.92       24.06
   05:bmi                               |         19        3.24       27.30
   06:tumor char                        |         16        2.73       30.03
   07:race                              |         16        2.73       32.76
   08:smoking                           |         15        2.56       35.32
   09:oc (oral contraceptives)          |         14        2.39       37.71
   10:pa (physical activity)            |         14        2.39       40.10
   11:diet                              |         11        1.88       41.98
   12:wt                                |         11        1.88       43.86
   13:SES                               |          9        1.54       45.39
   14:lactation hist                    |          9        1.54       46.93
   15:ht                                |          9        1.54       48.46
   16:educ                              |          7        1.19       49.66
   17:obesity                           |          7        1.19       50.85
   18:PCB                               |          6        1.02       51.88
   19:menopause                         |          6        1.02       52.90
   20:lactation                         |          6        1.02       53.92
   21:DDE                               |          6        1.02       54.95
   22:ER                                |          5        0.85       55.80
   23:mamm density                      |          5        0.85       56.66
   24:PR                                |          5        0.85       57.51
   25:dietary fat                       |          4        0.68       58.19
   26:screening                         |          4        0.68       58.87
   27:parenchymal pattern               |          4        0.68       59.56
   28:breast size                       |          4        0.68       60.24
   29:wt hist                           |          4        0.68       60.92
   30:BRCA1                             |          4        0.68       61.60
   31:abortion                          |          4        0.68       62.29
```

32:DDT	4	0.68	62.97
33:pregnancy	4	0.68	63.65
34:menstrual hist	4	0.68	64.33
35:hormones	3	0.51	64.85
36:body fat	3	0.51	65.36
37:hrt	3	0.51	65.87
38:estradiol	3	0.51	66.38
39:IGF	3	0.51	66.89
40:menarchial hist	3	0.51	67.41
41:occupation	3	0.51	67.92
42:Ashkenazi	3	0.51	68.43
43:BRCA2	2	0.34	68.77
144:p53	2	0.34	69.11
45:birthwt	2	0.34	69.45
46:ovarian ca	2	0.34	69.80
47:breastfeeding	2	0.34	70.14
48:antidepressants	2	0.34	70.48
149:fat	2	0.34	70.82
50:CYP (cytochrome P-450)	2	0.34	71.16
51:many	2	0.34	71.50
52:welldone meat	2	0.34	71.84
53:parental age	2	0.34	72.18
54:elec blanket	2	0.34	72.53
55:NAT2	2	0.34	72.87
56:ht hist	2	0.34	73.21
57:relative w brca	2	0.34	73.55
58:bmd	2	0.34	73.89
59:diet hist	2	0.34	74.23
60:NAT	2	0.34	74.57
61:cysts	2	0.34	74.91
62:ets (environ tobacco smoke)	2	0.34	75.26
63:maternal age	2	0.34	75.60
64:oc hist	2	0.34	75.94
65:radiation	2	0.34	76.28
66:her2neu	2	0.34	76.62
67:testosterone	2	0.34	76.96
68:breast density	2	0.34	77.30
69:child birth wt	1	0.17	77.47
70:breast cyst fluid	1	0.17	77.65
71:HCB (hexachlorobenzene)	1	0.17	77.82
72:hair dye	1	0.17	77.99
73:copper	1	0.17	78.16
74:cohort	1	0.17	78.33
75:biliary cirrhosis	1	0.17	78.50
76:urinary androgens	1	0.17	78.67
77:apoE	1	0.17	78.84
78:BRCA	1	0.17	79.01
79:urine melatonin	1	0.17	79.18
80:PBB (polybrominated biphenyls)	1	0.17	79.35
81:ATM (ataxia telangiectasia)	1	0.17	79.52
82:anthro	1	0.17	79.69
83:cholesterol	1	0.17	79.86
84:atyp hyper	1	0.17	80.03
85:allium veg	1	0.17	80.20
86:sebum	1	0.17	80.38
87:lactati	1	0.17	80.55
88:polio	1	0.17	80.72
89:insulin resistance	1	0.17	80.89
90:coping	1	0.17	81.06
91:progestens	1	0.17	81.23
92:His (sulfotransferase allele)	1	0.17	81.40
93:aspirin	1	0.17	81.57
94:pulse	1	0.17	81.74
95:pa hist	1	0.17	81.91
96:bilateral brca	1	0.17	82.08
97:heat shock proteins	1	0.17	82.25
98:NAT1	1	0.17	82.42
99:CYP3A4	1	0.17	82.59
100:GSTT1	1	0.17	82.76
101:dysplasia	1	0.17	82.94
102:COMT (catechol estrogen inact)	1	0.17	83.11

103:anthrop	1	0.17	83.28
104:twins zygosity	1	0.17	83.45
105:maternal breast feeding	1	0.17	83.62
106:erbB-2	1	0.17	83.79
107:tubal ligation	1	0.17	83.96
108:trigycerides	1	0.17	84.13
109:HDL	1	0.17	84.30
110:bbd	1	0.17	84.47
111:maternal hist	1	0.17	84.64
112:maternal cancer	1	0.17	84.81
113:death of partner	1	0.17	84.98
114:fibroadenoma	1	0.17	85.15
115:migration	1	0.17	85.32
116:dietary fiber	1	0.17	85.49
117:sexual assault	1	0.17	85.67
118:sterilization	1	0.17	85.84
119:hip fracture	1	0.17	86.01
120:albumin	1	0.17	86.18
121:K	1	0.17	86.35
122:homocysteine	1	0.17	86.52
123:condoms	1	0.17	86.69
124:SHBG	1	0.17	86.86
125:vitamins	1	0.17	87.03
126:adiposity	1	0.17	87.20
127:cholecystectomy	1	0.17	87.37
128:carotene	1	0.17	87.54
129:cytology	1	0.17	87.71
130:elec appliances	1	0.17	87.88
131:night employment	1	0.17	88.05
132:work exposure	1	0.17	88.23
133:menarche	1	0.17	88.40
134:demog	1	0.17	88.57
135:psychotrop med	1	0.17	88.74
136:CYP-450	1	0.17	88.91
137:parity	1	0.17	89.08
138:vitamin C	1	0.17	89.25
139:fat intake	1	0.17	89.42
140:lipids	1	0.17	89.59
141:birthmonth	1	0.17	89.76
142:fertility drugs	1	0.17	89.93
143:GTT	1	0.17	90.10
144:tissue removal	1	0.17	90.27
145:diabetes	1	0.17	90.44
146:Na	1	0.17	90.61
147:husband brca	1	0.17	90.78
148:TNFalpha	1	0.17	90.96
149:vit D	1	0.17	91.13
150:alcoholism	1	0.17	91.30
151:fiber	1	0.17	91.47
152:comorbidity	1	0.17	91.64
153:estrogen	1	0.17	91.81
154:geog	1	0.17	91.98
155:glucose	1	0.17	92.15
156:bp	1	0.17	92.32
157:farming	1	0.17	92.49
158:B12	1	0.17	92.66
159:breastfeeding hist	1	0.17	92.83
160:nsaids	1	0.17	93.00
161:time period	1	0.17	93.17
162:immigrants	1	0.17	93.34
163:induced abortion	1	0.17	93.52
164:B6	1	0.17	93.69
165:cell char	1	0.17	93.86
166:lefthandedness	1	0.17	94.03
167:GST	1	0.17	94.20
168:qol	1	0.17	94.37
169:treatment	1	0.17	94.54
170:comorbidities	1	0.17	94.71
171:GSTM1	1	0.17	94.88
172:breast reconstruction	1	0.17	95.05
173:folate	1	0.17	95.22

```
    174:remarriage                      |          1        0.17       95.39
    175:sunlight                        |          1        0.17       95.56
    176:various medical conditions      |          1        0.17       95.73
    177:multiple births                 |          1        0.17       95.90
    178:antibacterials                  |          1        0.17       96.08
    179:caffeine                        |          1        0.17       96.25
    180:occup emf                       |          1        0.17       96.42
    181:CYP17                           |          1        0.17       96.59
    182:FSH                             |          1        0.17       96.76
    183:selenium                        |          1        0.17       96.93
    184:HSD17B1                         |          1        0.17       97.10
    185:DHA                             |          1        0.17       97.27
    186:ascorbic acid                   |          1        0.17       97.44
    187:paternal age at birth           |          1        0.17       97.61
    188:surgery timing re menstrual cycle |        1        0.17       97.78
    189:preterm birth                   |          1        0.17       97.95
    190:psych hist                      |          1        0.17       98.12
    191:progesterone                    |          1        0.17       98.29
    192:ovarian pathology               |          1        0.17       98.46
    193:CD44 (transmembrane glycoprotein) |        1        0.17       98.63
    194:olive oil                       |          1        0.17       98.81
    195:atyp hyp                        |          1        0.17       98.98
    196:serum lipids                    |          1        0.17       99.15
    197:sex steroids                    |          1        0.17       99.32
    198:DMPA                            |          1        0.17       99.49
    199:hirsutism                       |          1        0.17       99.66
    200:familial clustering             |          1        0.17       99.83
    201:serum hormones                  |          1        0.17      100.00
--------------------------------------+-----------------------------------
                                Total |        586      100.00
```

Table 3 illustrates the astonishing diversity in the search for important risk factors for breast cancer. (Obviously, ranks above 42 are meaningless.) The counts here are of the numbers of articles in which a potential risk factor was investigated. Note that 132 of 201 factors (66%) were studied in only one article.

```
TABLE 4.
                                                 Freq.     Percent
--------------------------------------+---------------------------
Reproduction                                       111       18.94
   01:reprod hist                     |             50        8.53
   14:lactation hist                  |              9        1.54
   19:menopause                       |              6        1.02
   20:lactation                       |              6        1.02
   31:abortion                        |              4        0.68
   33:pregnancy                       |              4        0.68
   34:menstrual hist                  |              4        0.68
   35:hormones                        |              3        0.51
   37:hrt                             |              3        0.51
   38:estradiol                       |              3        0.51
   40:menarchial hist                 |              3        0.51
   47:breastfeeding                   |              2        0.34
   76:urinary androgens               |              1        0.17
   87:lactati                         |              1        0.17
   91:progestens                      |              1        0.17
  124:SHBG                            |              1        0.17
  133:menarche                        |              1        0.17
  137:parity                          |              1        0.17
  153:estrogen                        |              1        0.17
  159:breastfeeding hist              |              1        0.17
  163:induced abortion                |              1        0.17
  177:multiple births                 |              1        0.17
  189:preterm birth                   |              1        0.17
  191:progesterone                    |              1        0.17
  197:sex steroids                    |              1        0.17
  201:serum hormones                  |              1        0.17
```

```
----------------------------------------+-------------------------
Genetic                                       92         15.70
   02:family hist                       |     38          6.48
   07:race                              |     16          2.73
   30:BRCA1                             |      4          0.68
   42:Ashkenazi                         |      3          0.51
   43:BRCA2                             |      2          0.34
   144:p53                              |      2          0.34
   46:ovarian ca                        |      2          0.34
   50:CYP (cytochrome P-450)            |      2          0.34
   55:NAT2                              |      2          0.34
   57:relative w brca                   |      2          0.34
   60:NAT                               |      2          0.34
   66:her2neu                           |      2          0.34
   78:BRCA                              |      1          0.17
   92:His  (sulfotransferase allele)    |      1          0.17
   98:NAT1                              |      1          0.17
   99:CYP3A4                            |      1          0.17
  100:GSTT1                             |      1          0.17
  102:COMT (catechol estrogen inact)    |      1          0.17
  104:twins zygosity                    |      1          0.17
  106:erbB-2                            |      1          0.17
  111:maternal hist                     |      1          0.17
  112:maternal cancer                   |      1          0.17
  136:CYP-450                           |      1          0.17
  167:GST                               |      1          0.17
  171:GSTM1                             |      1          0.17
  181:CYP17                             |      1          0.17
  200:familial clustering               |      1          0.17
----------------------------------------+-------------------------
Behavioral                                    96         16.38
   04:alcohol                           |     23          3.92
   08:smoking                           |     15          2.56
   10:pa (physical activity)            |     14          2.39
   11:diet                              |     11          1.88
   25:dietary fat                       |      4          0.68
   26:screening                         |      4          0.68
   52:welldone meat                     |      2          0.34
   59:diet hist                         |      2          0.34
   85:allium veg                        |      1          0.17
   90:coping                            |      1          0.17
   95:pa hist                           |      1          0.17
  108:trigycerides                      |      1          0.17
  109:HDL                               |      1          0.17
  116:dietary fiber                     |      1          0.17
  125:vitamins                          |      1          0.17
  128:carotene                          |      1          0.17
  138:vitamin C                         |      1          0.17
  139:fat intake                        |      1          0.17
  140:lipids                            |      1          0.17
  149:vit D                             |      1          0.17
  150:alcoholism                        |      1          0.17
  151:fiber                             |      1          0.17
  158:B12                               |      1          0.17
  164:B6                                |      1          0.17
  173:folate                            |      1          0.17
  179:caffeine                          |      1          0.17
  183:selenium                          |      1          0.17
  186:ascorbic acid                     |      1          0.17
  194:olive oil                         |      1          0.17
----------------------------------------+-------------------------
Hazardous Exposure                            54          9.22
   09:oc (oral contraceptives)          |     14          2.39
   18:PCB                               |      6          1.02
   21:DDE                               |      6          1.02
   32:DDT                               |      4          0.68
   41:occupation                        |      3          0.51
   48:antidepressants                   |      2          0.34
   54:elec blanket                      |      2          0.34
   62:ets (environ tobacco smoke)       |      2          0.34
   64:oc hist                           |      2          0.34
```

```
     65:radiation                          |          2         0.34
     71:HCB (hexachlorobenzene)            |          1         0.17
     72:hair dye                           |          1         0.17
     73:copper                             |          1         0.17
     80:PBB (polybrominated biphenyls)     |          1         0.17
    130:elec appliances                    |          1         0.17
    131:night employment                   |          1         0.17
    132:work exposure                      |          1         0.17
    135:psychotrop med                     |          1         0.17
    142:fertility drugs                    |          1         0.17
    157:farming                            |          1         0.17
    180:occup emf                          |          1         0.17
---------------------------------------+-------------------------
Anthropometrical                                     65        11.09
     05:bmi                            |             19         3.24
     12:wt                             |             11         1.88
     15:ht                             |              9         1.54
     17:obesity                        |              7         1.19
     29:wt hist                        |              4         0.68
     36:body fat                       |              3         0.51
     45:birthwt                        |              2         0.34
     49:fat                            |              2         0.34
     56:ht hist                        |              2         0.34
     58:bmd                            |              2         0.34
     69:child birth wt                 |              1         0.17
     82:anthro                         |              1         0.17
    103:anthrop                        |              1         0.17
    126:adiposity                      |              1         0.17
---------------------------------------+-------------------------
Breast Physiology                                    22         3.75
     23:mamm density                   |              5         0.85
     27:parenchymal pattern            |              4         0.68
     28:breast size                    |              4         0.68
     61:cysts                          |              2         0.34
     68:breast density                 |              2         0.34
     70:breast cyst fluid              |              1         0.17
     84:atyp hyper                     |              1         0.17
    129:cytology                       |              1         0.17
    172:breast reconstruction          |              1         0.17
    195:atyp hyp                       |              1         0.17
---------------------------------------+-------------------------
Other Diseases/Conditions                            21         3.58
     75:biliary cirrhosis              |              1         0.17
     81:ATM (ataxia telangiectasia)    |              1         0.17
     88:polio                          |              1         0.17
     89:insulin resistance             |              1         0.17
    107:tubal ligation                 |              1         0.17
    110:bbd                            |              1         0.17
    114:fibroadenoma                   |              1         0.17
    117:sexual assault                 |              1         0.17
    118:sterilization                  |              1         0.17
    119:hip fracture                   |              1         0.17
    127:cholecystectomy                |              1         0.17
    143:GTT                            |              1         0.17
    144:tissue removal                 |              1         0.17
    145:diabetes                       |              1         0.17
    152:comorbidity                    |              1         0.17
    155:glucose                        |              1         0.17
    156:bp                             |              1         0.17
    170:comorbidities                  |              1         0.17
    188:surgery timing re menstrual cycle |           1         0.17
    190:psych hist                     |              1         0.17
    192:ovarian pathology              |              1         0.17
---------------------------------------+-------------------------
Other - hard to classify                            126        21.50
     03:age                            |             30         5.12
     06:tumor char                     |             16         2.73
     13:SES                            |              9         1.54
     16:educ                           |              7         1.19
     22:ER                             |              5         0.85
     24:PR                             |              5         0.85
```

39:IGF	3	0.51
51:many	2	0.34
53:parental age	2	0.34
63:maternal age	2	0.34
67:testosterone	2	0.34
74:cohort	1	0.17
76:urinary androgens	1	0.17
77:apoE	1	0.17
79:urine melatonin	1	0.17
83:cholesterol	1	0.17
86:sebum	1	0.17
93:aspirin	1	0.17
94:pulse	1	0.17
96:bilateral brca	1	0.17
97:heat shock proteins	1	0.17
101:dysplasia	1	0.17
105:maternal breast feeding	1	0.17
113:death of partner	1	0.17
115:migration	1	0.17
120:albumin	1	0.17
121:K	1	0.17
122:homocysteine	1	0.17
123:condoms	1	0.17
134:demog	1	0.17
141:birthmonth	1	0.17
146:Na	1	0.17
147:husband brca	1	0.17
148:TNFalpha	1	0.17
154:geog	1	0.17
160:nsaids	1	0.17
161:time period	1	0.17
162:immigrants	1	0.17
165:cell char	1	0.17
166:lefthandedness	1	0.17
168:qol	1	0.17
169:treatment	1	0.17
174:remarriage	1	0.17
175:sunlight	1	0.17
176:various medical conditions	1	0.17
178:antibacterials	1	0.17
182:FSH	1	0.17
184:HSD17B1	1	0.17
185:DHA	1	0.17
187:paternal age at birth	1	0.17
193:CD44 (transmembrane glycoprotein)	1	0.17
196:serum lipids	1	0.17
198:DMPA	1	0.17
199:hirsutism	1	0.17

Table 4 is an attempt to classify the risk factors from Table 3 into meaningful categories. This classification is obviously not the only one that could be used. It tends to be phenomenological, in the sense that, for example, many of the anthropometric measurements were intended to be indirect measures of endogenous estrogen exposure, but there is no estrogen exposure category. In other words, the categories were formed more for taxonomic than explanatory purposes. A summary appears in Table 5.

TABLE 5.

	Freq.	Percent
Reproduction	111	18.94
Genetic	92	15.70
Behavioral	96	16.38
Hazardous Exposure	54	9.22
Anthropometrical	65	11.09
Breast Physiology	22	3.75
Other Diseases/Conditions	21	3.58

```
Other - hard to classify*                         90       15.36
-----------------------------------------+-------------------------
*excluding "age" and "tumor char".
```

There is a fairly even split with respect to reproductive factors, genetic factors, and behavioral factors. As mentioned above, "anthropometric" factors are probably oriented toward endogenous estrogen exposure, irrespective of origin.

Discussion of Literature Results

The most prevalent (43%) article on breast cancer etiology is focused entirely on one risk factor as it relates to the disease. If breast cancer were a disease of one dominant cause, the scientific strategy implied by these studies would be justified. Since it is not, they are not.

Research on breast cancer is characterized by a concentrated investigation of a few generic risk factors (reproductive history, genetic susceptibility, behavioral aspects) with a variety of strategies for measuring these fundamental factors, and then a panoply of investigations of other types of risk factors that appears both opportunistic and undisciplined. A large number of articles seemed to have come from studies that were designed for some other purpose, with a breast cancer component attached as if an after-thought.

Although there is a large fraction (31%) of studies that go beyond the simplistic "one factor" model, for the most part each such study focuses on estimating the unique, independent contribution of a single risk factor of interest to breast cancer incidence or mortality. This view is diametrically opposed to the idea that the understanding of the causation of a disease involves comprehension of (1) how all of the risk factors are related in a causal system among themselves, and (2) which causal roles they play in producing the disease. For example, in none of the articles surveyed here was there an attempt to assess direct, indirect, or total causal effects, the minimal first step in a causal analysis. The concept of a minimal sufficient causal pathway appeared in no article.

Theoretical Results

Formally, modern causal inference reduces to the detection of independence and conditional independence conditions among a collection of chance variables that are assumed to be causally sufficient (no important causes have been left out). The notation that is used is $X \amalg Y|Z$ to mean that the chance variables in X are independent of the chance variables in Y, given the values of the chance variables in Z. The probability measure that is implicit in this involves random sampling from the operation of the causal system, and therefore does not pertain to retrospective sampling, which we have seen comprises the majority of breast cancer etiology studies. The usual retrospective sampling assumption is that the indicator of inclusion in the study is conditionally independent of all other variables, given the disease status. Under this condition the following result was proved: if the disease outcome variable is in X (or equivalently in

Y) or in Z, then the conditional independence relationship can be assessed in the retrospective study, and if the disease outcome does not appear among X, Y, or Z, then it is enough to know the marginal distribution of the disease outcome in the source population in order to assess the conditional independence condition.

While this result implies that a conventional causal analysis is not much more difficult from retrospective data than it is from prospective data, there is at least one critical additional problem. Conventional causal analysis does not deal well with temporal relationships. In effect, it assumes that all temporal processes have worked themselves out, leaving us with an accurate picture. The failure of modern causal analysis to deal effectively with time in prospective studies suggests that it will have an even harder time in retrospective studies.

In order to explore this aspect of the problem, it was necessary to stand back and take a synoptic look at temporal causal processes. For this purpose, the viewpoint was a combination of two approaches, *simulation* and *counter-factual causal theory*. The latter posits times at which events will occur as a consequence of causal processes. The former says that unless one understands enough to construct a valid simulation of a causal system, one does not yet understand it. Surprisingly, these two perspectives make it possible to develop and prove a number of results in temporal causal analysis. Specifically, a number of new methods and results were derived.

First, a method was developed for describing interdependent event times, based on the Möbius inversion theorem. From this, it is possible to compute marginal and conditional distributions of event times in a systematic manner. From the perspective of event simulation, it was discovered that only *local independence*, not full independence, was required to simulate event times as if they were fully independent. This is a new result that has wide implications in event simulation. It was also discovered, however, that the presumed marginal distributions that should be used in these simulations are not the marginal distributions of the event times. Examination of the Kaplan-Meier survival curve estimation procedure showed that the marginal distribution implied by this procedure is precisely what is required for simulation purposes. This is important because a number of authors have suggested recently that the K-M procedure does not estimate a biologically meaningful cumulative occurrence function, but the research completed under this project shows that this is not the case, and in fact the K-M estimate is precisely what is required for "independent" event simulation of dependent events.

This finding led naturally to the next question, is there any reason why the K-M procedure cannot be used on retrospective data? The fact that odds ratios are the same whether estimated retrospectively or prospectively lends some plausibility to this conjecture. Due to the successful characterization of joint event times, however, it was possible to show that K-M cannot be applied to retrospective data without inducing a bias, which was explicitly computed. Moreover, it was shown that even in retrospective situations there is a procedure based on a complementary exponential model that produces an unbiased estimate of the cumulative occurrence of disease. It was further shown that matching (on age, for example) makes it impossible to produce unbiased estimates. The best-known model for estimating a woman's risk of breast cancer, developed by Mitchell Gail at NCI, is based on an age-matched retrospective sample analyzed prospectively.

Discussion of Theoretical Results

Conventional causal analysis makes a number of fundamental assumptions, one of which is that all causal laws have played themselves out at the time we make our measurements on the causal system. In prospective studies one can, perhaps, make some allowances for the failure of such an assumption. In retrospective studies, however, it is necessary to take the timing of measurements into account in a way that is unfamiliar to epidemiologists, in order to make even the first few steps toward a causal analysis. For times to events, a general methodology for representing interdependent, counter-factual times has been developed, and it has been applied to prove results that are of immediate practical import. The importance of the K-M method has been reaffirmed, although on a different basis (simulation and causation) than is generally understood in the literature.

Key Research Accomplishments

- Literature review of recent breast cancer etiology studies
- Development of method for classifying inferential structures
- Tabulation of risk factors by their intensity of study
- Finding a lack of causal analysis in breast cancer etiology
- Development of a new general method for representing dependent event times
- Determination that Kaplan-Meier approach is appropriate for simulation, despite recent research to the contrary
- Determination that Kaplan-Meier approach is not appropriate for retrospective studies; computation of exact bias
- Development of a valid complementary exponential model for retrospective studies with time-to-event data
- Determination that the conditional independence conditions of modern causal analysis can be tested with retrospective data (although in one case prevalence data is require); but the atemporal nature of modern causal analysis does not easily apply to time-to-event data in retrospective studies

Reportable Outcomes

Article on simulation and causation with new results on time-to-event analysis, particularly in regard to retrospective studies (in prep)

Article on the structure of the recent breast cancer etiology literature (in prep)

Conclusions

Discussion of Literature and Theoretical Results

The prevalence of retrospective studies suggests that most of the information on breast cancer etiology that we are likely to acquire will come from this kind of study. At a gross level of approximation, it has been shown in this project that the aims of modern causal analysis (detection of independence and conditional independence) are as feasible from retrospective data as they are from prospective data. Closer consideration suggests, however, that a more sophisticated analysis of the timing of events might represent a major step forward in understanding breast cancer etiology, and that provided the appropriate measurements are made, this information is also as available in retrospective studies as it is in prospective studies, even though the methods of extracting it are new.

The fundamental difficulties in breast cancer etiology are illustrated by the very wide variety of risk factors that have been investigated. There have been no attempts to bind these (mostly) single-factor studies into anything like a causal web for breast cancer etiology. This project has succeeded in developing some tools that might, with appropriate data, begin to make such an enterprise think-able, even if most of the data were to come from retrospective studies. The implication is, however, that these datasets would have to be collected into a single archive, in order to make the interconnections between them that are necessary for a *causal simulation* approach to breast cancer.

The "so what" result is as follows. Breast cancer is a disease of highly multifactorial etiology, so far as we can tell, and based on a literature review. Inferential methodology in breast cancer research is structured as if the disease had a few major causes. Much of the existing research is retrospective in nature. Even though modern causal analysis is designed for prospective studies, the retrospective studies can still contribute provided (1) different time-to-event methods are used than are used in prospective studies, and (2) raw datasets are available for re-analysis.